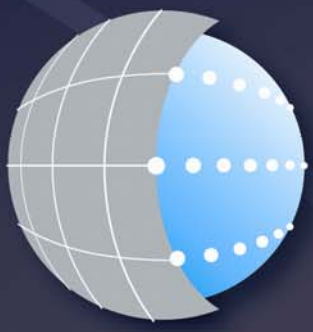
National Cancer Institute



caBIG™

cancer Biomedical Informatics Grid ™

caBIG™ Primer:
An Introduction to caBIG™

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

December 2006

The cancer Biomedical Informatics Grid™ (caBIG™) Primer

TABLE OF CONTENTS

Document Overview

Why a caBIG™ Primer?

The caBIG™ Primer provides a high level overview of caBIG™, the cancer Biomedical Informatics Grid™. The audience for this Primer includes current, new, and prospective caBIG™ participants and beneficiaries who desire an overview of the program's vision and mission, organization, activities, and challenges.

The intent is to provide an overview that gives a "lay of the land" supplemented by a list of next steps that allow readers to pursue specific areas driven by interest and organizational goals.

Approach to this Document

This Primer is organized into several sections, covering the following topics:

- The Need and Vision for caBIG™
- Mission, Principles and Initiation of caBIG™
- Organizing caBIG™: Operational Setting
- Achieving Interoperability
- Opportunities to Contribute and Benefit in caBIG™

The Need and Vision for caBIG™

Defining the Challenge

caBIG™ is sponsored by the National Cancer Institute (NCI), and is administered by the National Cancer Institute Center for Bioinformatics (NCICB). Dr. Andrew von Eschenbach, former Director of the National Cancer Institute (NCI) and Dr. Kenneth Buetow, Director, NCI Center for Bioinformatics, have clearly expressed the opportunities and challenges leading to caBIG™:

"We are in the midst of an explosion of knowledge about cancer as a disease process. We are beginning to understand cancer not by what we can see and touch – or by what is revealed under a microscope – but at the molecular level. It is not a question of if, but rather when and how, molecular medicine translates into personalized care…

"We cannot achieve this (translation) without greater interconnectivity and coordination across the cancer enterprise. This requires seeing cancer as a systems problem that will require a systems solution. Although cancer is being unraveled rapidly at the genomic and proteomic levels, we have not concomitantly developed the seamless system needed to capitalize on our discoveries.

"To universally integrate personalized medicine into cancer prevention, diagnosis and treatment, researchers and clinicians must be able to gain rapid access to multiple types of specific information about an individual patient -- information to which they do not currently have easy access. A new generation of medicine will require incorporation of shared information technologies (IT). NCI is committed to creating a standards-based biomedical infrastructure and to making it available across the entire cancer enterprise."[1]

Addressing these challenges involves several tasks:

- Analyzing and integrating vast amounts of data from disparate sources.
- Sharing tools for broader application.
- Utilizing and developing new standards where appropriate and a unifying infrastructure to enable the sharing of data and resources across the cancer research community.

caBIG™ was founded to help maximize the full power of our collective knowledge about cancer, maximizing opportunities for both scientific discovery and patient care.

The caBIG ™ Vision

The vision of caBIG™ is a full cycle of integrated cancer research, extending from bench to bedside, and back again. This requires launching "an international collaboration to facilitate and enable research teams to share data, applications, and infrastructure…."[2]

Sharing and integrating cancer knowledge through a common infrastructure is the new paradigm envisioned by caBIG™. The affiliation of caBIG™ participants with a broad range of expertise promises to accelerate progress in all aspects of cancer research - including cancer origin, prevention, early detection and treatment.

caBIG™ will help redefine how cancer research is conducted and eventually, how cancer care is provided.

[1] Andrew C. von Eschenbach, M.D. and Kenneth Buetow, Ph.D. (2006) "Cancer Informatics Vision: caBIG™" Cancer Informatics 2006:2 (22-24)
[2] *Ibid.*

Mission, Principles, and Initiation of caBIG™

The Mission of caBIG™

The mission of caBIG™ is to provide infrastructure for creating, communicating and sharing bioinformatics tools, data and research results, using shared data standards and shared data models. This supports the development of new types of analysis within and across experiments and allows new forms of collaboration, enabling the sharing of data sets and a range of analytical tools. Figure 1 illustrates these concepts.

caBIG™ encourages researchers to focus on new endeavors, such as:

- Analyzing and integrating the vast amounts of information generated in the areas of genomics and proteomics.
- Identifying and applying information from clinically and molecularly annotated biospecimens in research data.
- Integrating information from a wide range of sources, in support of translational and personalized medicine.
- Using outcome information from relevant clinical trials to inform decision making.

The caBIG™ initiative will produce these results:

- A shared interoperable infrastructure within which researchers can collaborate.
- A set of common data elements, data models, and standards to better enable data sharing and integration.
- Tools for analyzing information associated with cancer research and care

"With caBIG™, the research community can focus its attention on innovation in prevention, diagnosis and treatment, not on data management and constantly building and troubleshooting the basic underlying research infrastructure." (Ken Buetow, PhD, NCI Associate Director for Bioinformatics and Information Technology)

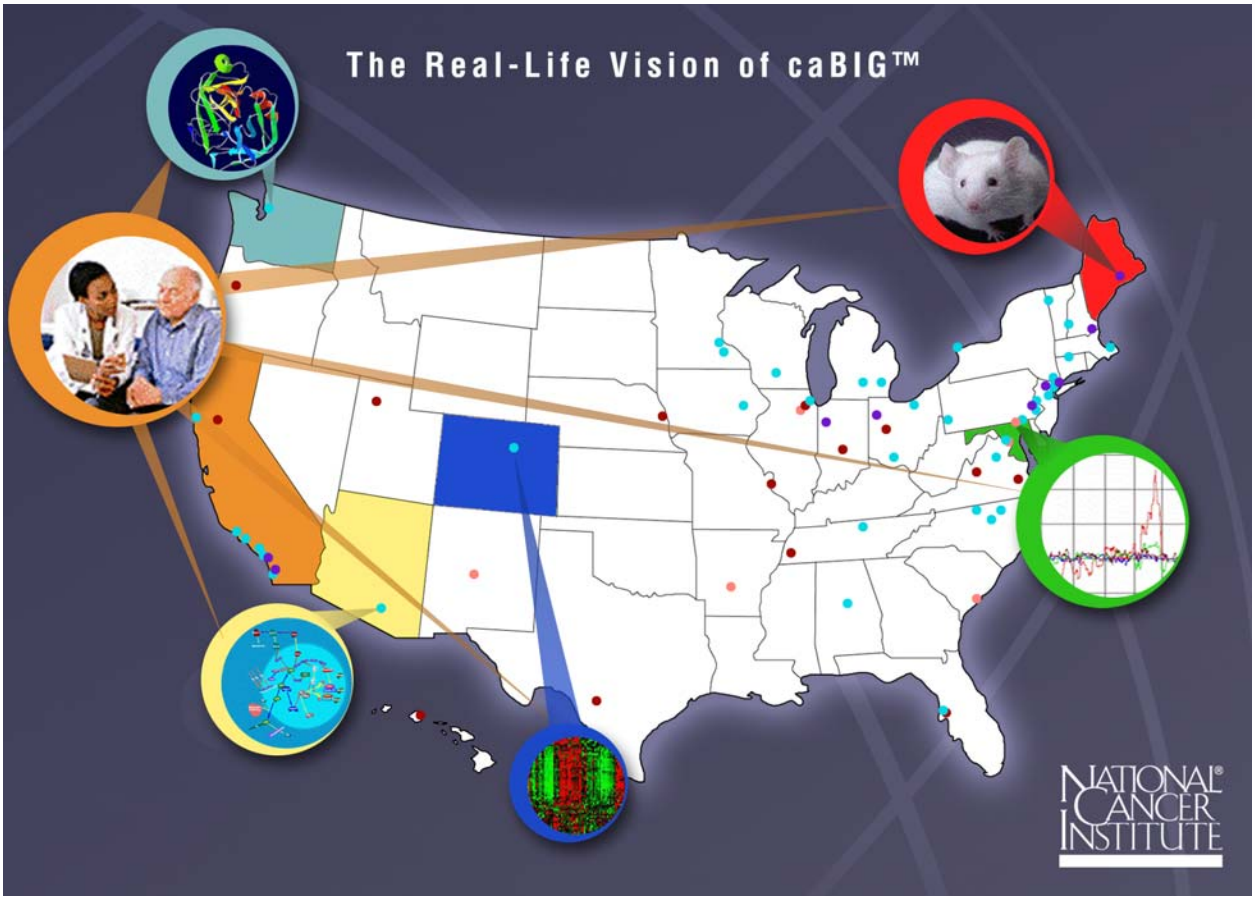


Figure 1 This real-life scenario illustrates the vision of caBIG™

Figure 1. The caBIG™ infrastructure enables the sharing of tools and data, and the integration of diverse data types across the cancer research enterprise. Drawing on expertise nationwide, caBIG™ will facilitate the flow of research results to support patient treatment and care, and the flow of treatment response data back to the laboratory bench. Together, this collaboration will accelerate the patient-centric molecular medicine revolution.

caBIG™ Principles

caBIG™ is based upon four fundamental principles:

- **Open Access**: Participation in caBIG™ and the products delivered by caBIG™ are open to all, enabling access to tools, data, and infrastructure by the cancer and greater biomedical research community.

- **Open Development**: Software development projects are assigned to particular participants, but are carried out iteratively with multiple opportunities for review, comment, further modification and development.

- **Open Source:** The software code underlying caBIG™ tools is available to software developers for use and modification. Software funded through caBIG™ is licensed as open source to promote the reuse of existing code, hence optimizing the full benefit of the research dollars spent. However, the open source license is industry-friendly, fostering industry interest and innovation, while still adhering to the principle of open source for caBIG™-funded activities.

- **Federation:** caBIG™ software and standards enable local organizations, such as cancer centers, to share computing or data resources with the larger cancer care and research community, and to use resources contributed by others. Within the grid, these resources can be aggregated from multiple sites to appear as an integrated research tool set or large database, while the individual resources remain under the control of the local organizations. This strategy of organizing and providing distributed access to locally-managed tools and data is referred to as “federation” and represents an alternative to centralized large-scale repositories and systems.

The Initiation of caBIG™

The first step in achieving the vision of caBIG™ was an outreach effort to NCI-designated cancer centers to identify their strengths, capabilities, and needs.

The most pressing and shared needs across cancer centers, illustrated in Figure 2, shaped the creation of caBIG™ workspaces. For example, the need for clinical data management tools led to the creation of the Clinical Trials Management Systems Workspace. These workspaces, or areas of focus, became the key organizational units for the caBIG™ pilot, launched in February 2004 under the coordinating supervision of the NCICB. Initially, the three year pilot was a collaboration between NCI and NCI-designated cancer centers; additional collaborations with government, academia, industry, and international groups were established later.

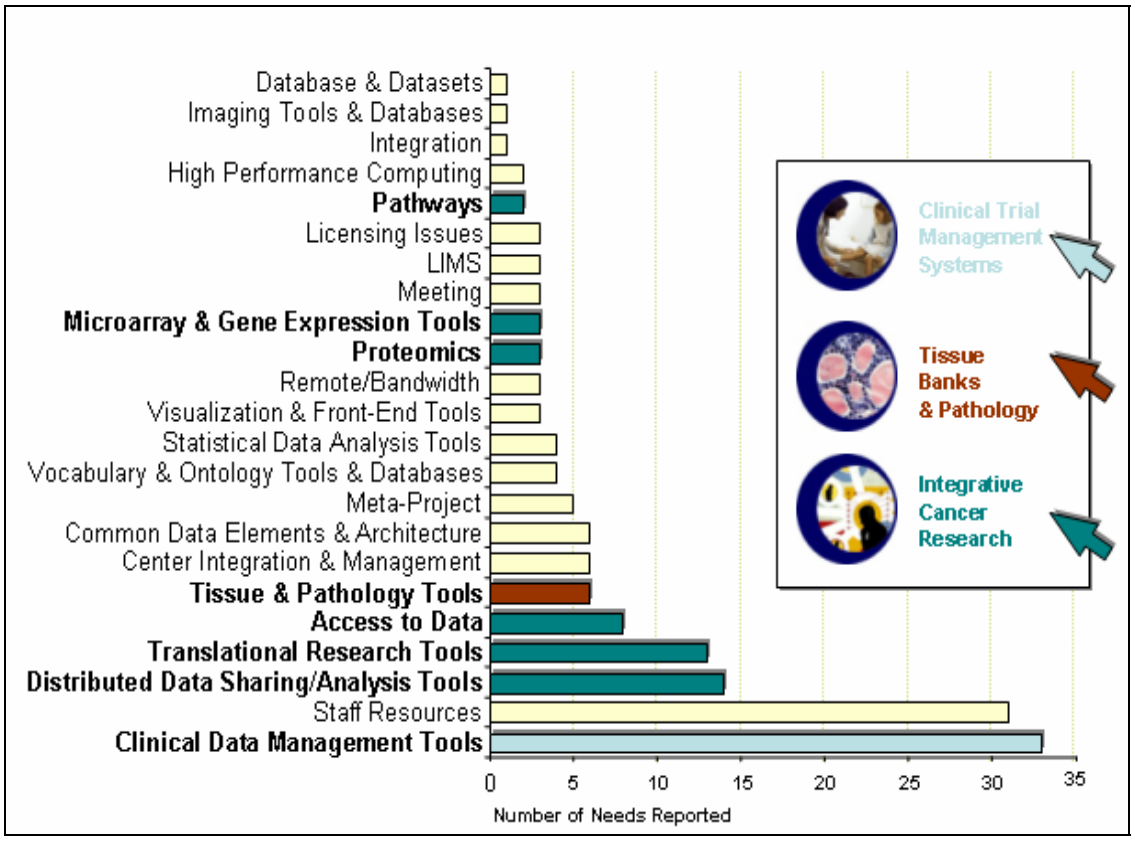


Figure 2 Cancer center involvement identified priority areas for caBIG™

caBIG™ Participants

caBIG™'s eventual goal is to involve the entire cancer community, encompassing the following groups:

- Those who perform basic science research on the origins and mechanisms of cancer (etiologic research).
- Those who study prevention, early detection, and treatment
- Those involved in drug development.
- Those who conduct clinical trials to bring effective new diagnostics and treatments to patients.
- Pathologists and oncologists at local hospitals involved in a patient's diagnosis and treatment.
- Clinicians and patients who can benefit from the seamless flow of information from bench to bedside.

In Year 3 of its pilot, caBIG™ participants include over 900 individuals from approximately 100 institutions. More than 50 NCI-designated cancer centers and roughly 10 additional institutions have executed base agreements, agreeing to share tools, data and common infrastructure to agreed-upon common standards. For those new to caBIG™, key players involved in the caBIG™ pilot are described below.

The National Cancer Institute Center for Bioinformatics (NCICB) – The mission of the NCICB is to provide foundational biomedical informatics infrastructure, tools and data to serve NCI research initiatives and the cancer research community. NCICB facilitates the integration of diverse research and data across the "bench to bedside continuum" through an interoperable infrastructure and collaborative leadership.

NCI-designated Cancer Centers – caBIG™ was conceived and developed with the input of NCI-designated cancer centers. Teams from the cancer centers and volunteers from within the cancer community continue to shape caBIG™ priorities by developing, refining, and testing tools; providing data sets; and setting policy and guidelines.

Patient Advocates – Patient advocates have been active members of the caBIG™ community since its inception. Each workspace includes a patient advocate "to help ensure that the caBIG™ end product, the Grid, will ultimately benefit the cancer patient by improving patient care and outcomes in the most effective and timely way possible "[3]

Industry Partners – Currently, many commercial organizations and some not-for-profit organizations participate in caBIG™ activities. These organizations include information technology companies, large-scale software vendors, pharmaceutical companies, and biotechnology companies as well as small, specialized ventures. Participation varies from voluntary involvement in workspace teleconferences and face-to-face meetings to funded development of caBIG™ applications.

[3] Statement of Expectations, Purpose and Goals From the caBIG™ Patient Advocates - caBIG website.

Additional participants include other **federal agencies, academic centers**, and **members of the international community**.

NCI welcomes new adopters of and participants in the caBIG™ program. In fact, given the applicability of caBIG™'s foundational infrastructure and tools, it is anticipated that caBIG™ will readily accommodate not only the broader cancer research community, but the biomedical research community at large.

Organizing caBIG™: Operational Setting

Organizing caBIG™

Participants in caBIG™ fulfill one or more roles:

- Developing or modifying interoperable tools (e.g., software, infrastructure).
- Adopting applications for use in settings different from those in which they were developed.
- Mentoring others in data model and tool development, software development, documentation or training activities.
- Contributing to white papers in strategic, policy or technology areas, such as patient privacy or security architecture.

These activities are carried out primarily through caBIG™ workspaces. The initiative is actively managed by a general contractor and the National Cancer Institute to ensure that priorities are met in a coordinated and timely fashion.

Workspaces – caBIG™ activities take place within workspaces. Participants provide data, subject matter expertise, and solutions; create and evaluate standards; and help shape the workspace's strategic direction – ultimately identifying, prioritizing, and fulfilling the needs of the specific area of focus of the workspace.

In addition to contributing to ongoing programmatic workspace activities, caBIG™ participants may also be part of developer/adopter project teams, focused on specific tools or other needed products. Developers construct tools or adapt existing tools to address needs identified by the cancer community. Adopters provide data sets, and focus on testing, validating, and applying tools developed by others for their own institution's workflow.

Workspaces are categorized as:

- Domain Workspaces (focused on specific disciplines)
- Cross-Cutting Workspaces (focused on defining and achieving interoperability)
- Strategic-Level Workspaces (focused on overarching issues integral to all workspaces)

Special Interest Groups (SIGs) within workspaces support these groups, and are convened to address specific or specialized needs of workspaces as they arise. Collectively, the

workspaces are not only building the foundation for caBIG™, they are also defining and refining caBIG™'s goals, priorities and activities.

Addressing Identified Needs: The Domain Workspaces

The Domain Workspaces were formed to respond to needs identified by NCI-sponsored cancer research centers:

- **Clinical Trial Management Systems:** Develops a comprehensive set of modular, interoperable and standards-based tools designed to meet the diverse clinical trials management needs of the Cancer Center community.
- **Integrative Cancer Research Workspace:** Produces modular and interoperable tools and interfaces that provide for integration between biomedical informatics applications and data. This will ultimately enable translational and integrative research by providing for the integration of clinical and basic research data.
- ***In Vivo* Imaging Workspace:** Creates, optimizes and validates tools and methods to extract meaning from and share imaging data.
- **Tissue Banks and Pathology Tools Workspace:** Develops a set of tools to inventory, track, mine, and visualize biospecimens and related annotations from geographically dispersed repositories.

Setting Standards: The Cross-Cutting Workspaces

The Cross-Cutting Workspaces create or specify the standards and infrastructure used by all Domain Workspaces to insure interoperability as they develop their respective tools and systems. These workspaces also support mentoring, and guide and support work related to reference implementations:

- **Architecture Workspace**: Develops the platform that integrates diverse data types and supports interoperable analytic tools. This group also defines syntactic interoperability, mentors developers, and is developing caGrid. caGRID is the underlying network architecture and platform that provides the basis for connectivity, tools deployment, and data sharing between caBIG™ participants.
- **Vocabularies and Common Data Elements Workspace**: Evaluates and integrates systems and standards for vocabularies and common data elements and ontology content development, as well as software systems for content delivery. They also define semantic interoperability, train and provide mentors, and provide guidelines for the adoption of standards and CDE harmonization.

Creating Policies: The Strategic Level Workspaces

The Strategic Level Workspaces develop policies and guidelines that support other workspaces, and develop and refine the caBIG™ strategic plan:

- **Data Sharing and Intellectual Capital** - Addresses issues and develops recommendations related to data sharing, patient privacy, intellectual capital, security and other policies related to caGRID as well as other regulatory and proprietary issues
- **Documentation and Training** - Defines guidelines, processes, templates and tools for developing consistent software documentation and training materials and for fostering mentoring activities throughout caBIG™.

- **Strategic Planning** - Assists caBIG™ senior leadership with strategic planning and vision development activities.

Achieving Interoperability

Describing Interoperability

Interoperability is the ability of two or more systems to exchange information and to use the data that have been exchanged[4]. This definition signifies that there are actually two parts to interoperability: 1) accessing information from a system ('syntactic interoperability') and 2) using or understanding the data that has been received ('semantic interoperability'). To be truly interoperable, systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible, and the data itself can be translated by any of the systems into an understandable form.

From its inception, caBIG™ considered interoperability an essential component for meeting its mission. To this end, the program established caBIG™ compatibility guidelines to assist in the development of interoperable systems. The guidelines provide standards for interoperable systems in four areas:

Common Application Programming Interface Integration: The syntactic part of interoperability. The guidelines in this area provide standards for access to an electronic data system.

Vocabularies/Terminologies and Ontologies: A part of semantic interoperability. These provide guidance about the types of controlled terminology that are used to record information in the system and about the system.

Data Elements: A part of semantic interoperability. Data elements provide a detailed description of the meaning of the information that is recorded, in addition to its value (often called semantic metadata). For example, if an electronic system maintains information about a patient's temperature, the semantic metadata describes what is meant by a 'patient' and 'temperature' and what constitutes a valid value for a patient

[4] IEEE – Institute of Electrical and Electronics Engineers

temperature (probably a number between 0 and 100, measured in degrees Fahrenheit).

Information Models: **A part of semantic interoperability. An information model describes the structure of the data maintained in the grid system.**

Levels of caBIG™ Compatibility

Differing degrees of interoperability can be qualified in terms of the tool's adherence to the caBIG™ Compatibility Guidelines. Four different levels of maturity are defined: Legacy, Bronze, Silver, and Gold, with each maturity level reflecting a different level of syntactic and semantic interoperability.

Projects funded by caBIG™ are required to be compatible at silver level (at a minimum), which achieves both syntactic and semantic interoperability. The gold level is equivalent to grid-enabled and allows a group to advertise their service on the grid and to discover and invoke services, resulting in the applications being seamlessly integrated at the electronic level.

For more information about caBIG™Compatibility Guidelines, refer to the following web page: https://cabig.nci.nih.gov/guidelines_documentation/.

Opportunities to Contribute and Benefit in caBIG™

The Expanding caBIG™ Community

As the caBIG™ program and infrastructure continue to grow, more and more organizations are becoming involved. caBIG™ will increasingly involve a cross-section of diverse users, contributors and beneficiaries.

- End-users, who want to access or adopt existing tools.
- Developers and legacy system managers, who want to learn the cost-benefit of developing or modifying tools or systems that interoperate with the caBIG™ infrastructure.
- Clinicians wishing to set up a clinical trial management system that is compatible with caBIG™ standards, allowing for potential sharing and integration of results.
- Community hospitals that want to use caBIG™ tools to support patient care.
- Industry partners who want to develop and/or adopt caBIG™ tools.

NCICB, in collaboration with caBIG™ participants, continues to provide a series of enabling technologies and processes to support developers wishing to build caBIG™-compatible data systems. The open development policy established by caBIG™ provides reusable software to development groups. Processes are ongoing to create well-defined data standards.

caBIG™ Tools - Development and Distribution

Development and Status – The current inventory of caBIG™ tools, infrastructure and datasets can be accessed from the "Inventory Of Tools" section of the caBIG™ website, https://caBIG.nci.nih.gov/inventory. This inventory includes key infrastructure, applications and data sets from the NCICB and collaborating caBIG™ participants. They are available at different phases of deployment and interoperable readiness.

Utilizing caBIG™ tools - To adopt a software tool offered by the caBIG™ community, a potential adopter should first identify the tool on the caBIG™ website. In some cases, the tool and its documentation can be downloaded to begin using it. In other cases, tools listed on the caBIG™ website are currently under development. One should follow the appropriate links to learn more about the resources and information for adopting caBIG™ tools. Mentors, applications support and training resources are available to help new users as they evaluate and explore new tools. Working with the workspace participants and leadership in areas of interest can also be a means of quickly learning about the tools, infrastructure and data available to researchers.

Contribution of caBIG™ Candidate Tool or Data Set - Researchers or organizations with tools and data sets of potential use by the wider community can contact the appropriate workspace lead to determine specific next steps. Resources and informational tools will continue to be provided by the caBIG™ community and NCICB to assist those who want to contribute to the network.

Continuing Opportunities to Learn and Participate

A variety of ways exists to learn about and become involved with the caBIG™ community. All events and materials are available through the caBIG™ website: https://cabig.nci.nih.gov/, and many are also distributed through the caBIG™ LISTSERVS described below.

caBIG™ Website – The caBIG™ website offers overview materials about the caBIG™ project, including introductory slides and archived videocast seminars, information on primary caBIG™ resources and tools, a list of current participants, contact information, workspace minutes and documents, and much more. This information can be accessed at: https://caBIG.nci.nih.gov.

LISTSERVS - Many caBIG™ workspaces and project teams maintain LISTSERVS to announce upcoming events and new products. Refer to https://list.nih.gov/archives/cabig_announce.html to sign up for the general announcement list for caBIG™. Other listservs can be accessed through https://list.nih.gov. Instructions for subscribing are on this entry page.

Program Updates – Periodic program updates provide an in-depth look at key areas of caBIG™, and highlight the successes of the growing caBIG™ community. The program updates are available through the caBIG™ website, and are sent through the general announcement listserv.

Workspace Teleconferences - One way to get a sense of the caBIG™ project, and to help shape its direction, is to identify a workspace of interest and then request access to the next teleconference. A list of the workspaces and the primary contacts for each of the workspaces are available on the caBIG™ Website and through "What's BIG This Week". The workspace leads will provide you with the toll free dial-in number and pass code for a teleconference. Teleconferences take place on a regular basis.

Face-to-Face Meetings - Workspaces gather for face-to-face meetings periodically over the course of the year. These meetings provide the opportunity for members to tackle specific issues that are best accomplished together in real time; receive updates; and discuss accomplishments, upcoming tasks and future goals.

Town Hall Meetings - Periodically, the caBIG™ community gathers for a teleconference "town hall meeting", designed to provide updates and to address issues or questions about caBIG™. Prior to the Town Hall, questions are solicited from the community. Watch for announcements on the caBIG™ Announce listserv.

Annual Meeting - The caBIG™ community gathers annually to celebrate successes; address caBIG™-wide and cross-workspace topics; and demonstrate emerging and available products including software tools, databases, prototypes, white papers, and development models. Interested individuals are welcome at this meeting, with special sessions offered for newcomers.

Contact with the caBIG™ Team – The caBIG™ website contact page lists the leads and contact information for all caBIG™ workspaces: https://caBIG.nci.nih.gov/contact_us. For help with caBIG™ tools, contact NCICB applications support at ncicb@pop.nci.nih.gov; Telephone: 301-451-4384; Toll free: 888-478-4423

Funding Process and Opportunities

Funding for products and specific participant work is awarded through openly-competed contracts, driven by priorities established in the workspaces and the caBIG™ strategic plan. Funding opportunities are announced on the front page of the caBIG™ website and distributed through the caBIG™ Announce Listserv.

Looking Ahead: Facing Challenges Together

The vision and need for caBIG™ have been established, and tremendous achievements made. Despite this, challenges remain:

- Educating the broader community about the available tools and their suitability for a particular research or clinical need.
- Creating methods and a useful workflow for data sharing between basic science and clinical research.
- Reconciling data models and terminology across traditionally separate areas of science and clinical practice.
- Addressing issues of data sharing, including ensuring data security, patient privacy, and intellectual capital.
- Marshalling the diverse resources and standards developed in a variety of settings, and organizing them to address the common problem of cancer research and care.
- Coordinating and correctly sequencing multiple interdependent software and standards development projects to meet broader community needs.

As with any innovative and comprehensive initiative, the challenges of caBIG™ provide the opportunity for collaborative solutions that, in time, will reshape the overall cancer research and treatment paradigm.

Acknowledgements

The caBIG™ Primer was created by the caBIG™ Documentation and Training Workspace. The following contributors dedicated extensive time and efforts in the drafting of this document, and deserve much appreciation and recognition for their work (listed in alphabetical order):

- Mary Jo Deering - NCICB
- Leslie Derr - NCICB
- Jill Hadfield - NCICB
- Jim Harrison – University of Virginia
- Jamie Keller – NCICB/TerpSys
- Valerie Monaco - University of Pittsburgh
- Parul Purohit – University of California - Davis
- Todd Scheetz - University of Iowa – Holden
- Jennifer Tucker – OKA (Otto Kroeger Associates)